Learning actionness via long-range temporal order verification

Supplementary material

Dimitri Zhukov[1,2], Jean-Baptiste Alayrac[3], Ivan Laptev[1,2], and Josef Sivic[1,2,4]

[1] Département dinformatique de lENS, ENS, CNRS, PSL University, Paris, France
[2] INRIA, Paris, France
[3] DeepMind
[4] CIIRC, Prague, Czech Republic

1 Outline

This supplementary material presents additional quantitative and qualitative results of our method. In Section 2 we provide the results of Action vs. Background classification for each domain in the COIN dataset. In Section 3 we show background clips with the highest scores and action clips with the lowest scores from the COIN dataset.

2 Per-domain results

Figure 1 shows the average precision of frame-wise Action vs. Background classification for every domain in the COIN dataset. We compare our results to the Chance baseline. These results are obtained in the same setup as described in Section 4.1 of the main paper. As shown in 1, our model outperforms chance for every individual domain. The gap between our method and chance ranges from 9% for Sport to up to 26% for Drink and Snack. This shows that our method allows to separate actions from background in diverse instructional videos.

3 False positives and false negatives

Figure 2 illustrates the highest scoring COIN clips according to our model that do not intersect with any action segments in COIN annotation, and the lowest scoring COIN clips that lie within action segments in COIN annotation. Interestingly, all the highest scoring background clips in the top row of Figure 2 contain valid actions, that are not included in the COIN annotation.

Finally, when looking at the clips that score low according to our model but are labeled as action in the COIN dataset (bottom row of Figure 2), we see that only two clips actually depict an action: clip 6 (reveal the glue from the face) and clip 8 (scratch the hair carefully). The rest of the clips dont contain any action and are labeled as action due to imprecise annotation of action boundaries.

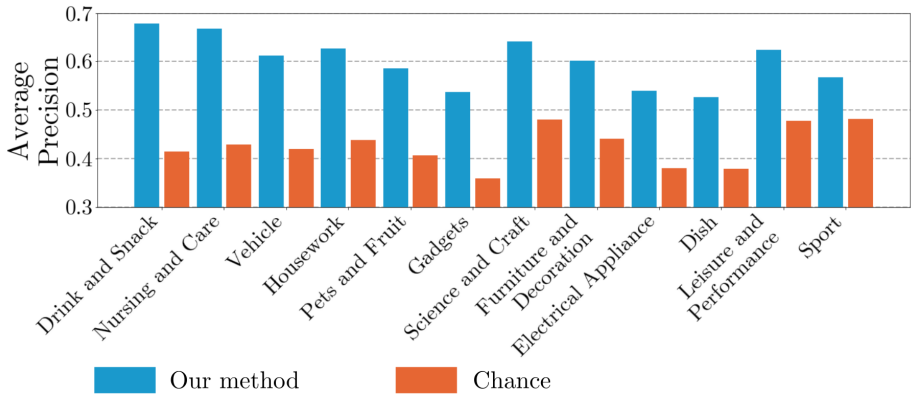

Fig. 1. Average precision of frame-wise Action vs. Background classification for every COIN domain. Our method outperforms chance in every individual domain

Fig. 2. Top: Highest scoring COIN clips, which are not annotated as action. **Bottom:** Lowest scoring COIN clips, which are annotated as action